RESEARCH-ARTICLE

FAME: Fusion-Aware Multi-modal Ensemble for Social Media Popularity Prediction

YAN ZHUANG, University of Electronic Science and Technology of China, Chengdu, Sichuan, China

WEI BAI, Tsinghua University, Beijing, China

YANRU ZHANG, University of Electronic Science and Technology of China, Chengdu, Sichuan, China

MINHAO LIU, University of Electronic Science and Technology of China, Chengdu, Sichuan, China

JIAWEN DENG, University of Electronic Science and Technology of China, Chengdu, Sichuan, China

FUJI REN, University of Electronic Science and Technology of China, Chengdu, Sichuan, China

Published: 27 October 2025

Citation in BibTeX format

MM '25: The 33rd ACM International Conference on Multimedia
October 27 - 31, 2025
Dublin, Ireland

Conference Sponsors:
SIGMM

https://doi.org/10.1145/3746027.3763759
ISBN: 9798400720352

FAME: Fusion-Aware Multi-modal Ensemble for Social Media Popularity Prediction

Yan Zhuang
University of Electronic Science and Technology of China
Chengdu, China
202211081370@std.uestc.edu.cn

Wei Bai
Tsinghua University
Beijing, China
bw23@mails.tsinghua.edu.cn

Yanru Zhang
University of Electronic Science and Technology of China
Chengdu, China
Shenzhen Institute for Advanced Study, UESTC
Shenzhen, China
yanruzhang@uestc.edu.cn

Minhao Liu
University of Electronic Science and Technology of China
Chengdu, China
Shenzhen Institute for Advanced Study, UESTC
Shenzhen, China
minhaoliu@uestc.edu.cn

Jiawen Deng*
University of Electronic Science and Technology of China
Chengdu, China
dengjw@uestc.edu.cn

Fuji Ren*
University of Electronic Science and Technology of China
Chengdu, China
Shenzhen Institute for Advanced Study, UESTC
Shenzhen, China
renfuji@uestc.edu.cn

Abstract

As social media becomes a dominant platform for sharing content, predicting the popularity of user posts has become increasingly important for applications such as content recommendation, trend forecasting, and user engagement. However, this task is challenging due to the diverse and multimodal nature of social media posts, which often include unstructured text, images, and structured metadata. To address this challenge, we propose **Fusion-Aware Multi-modal Ensemble (FAME)**, a framework effectively captures and integrates diverse information sources within social media content. Unlike prior approaches that rely on a single model to process all modalities, FAME leverages four specialized predictors. Three of them—CatBoost, LightGBM, and AutoGluon—are tree-based models that excel at handling structured metadata and its interactions with unstructured features. The fourth is a denoising autoencoder (DAE), which learns robust joint representations from unstructured text and image data. These models are combined through a weighted ensemble strategy, allowing FAME to leverage the complementary strengths of different architectures. Experiments on the Social Media Prediction Dataset demonstrate that FAME significantly outperforms existing baselines, achieving state-of-the-art results and validating its effectiveness in modeling the complex, multimodal nature of social media content.

*Corresponding authors.

MM '25, Dublin, Ireland

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3763759

CCS Concepts

• **Information systems** → **Web searching and information discovery**; • **Computing methodologies** → **Natural language processing**; **Computer vision**.

Keywords

Social Multimedia; Popularity Prediction; Multimodal Learning

ACM Reference Format:
Yan Zhuang, Wei Bai, Yanru Zhang, Minhao Liu, Jiawen Deng, and Fuji Ren. 2025. FAME: Fusion-Aware Multi-modal Ensemble for Social Media Popularity Prediction. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3746027.3763759

1 Introduction

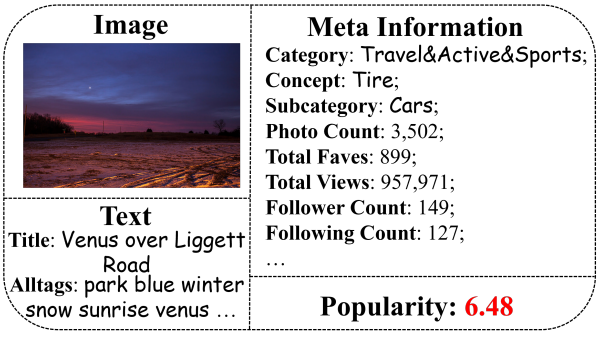

Figure 1: An example social media post from Flickr, including multimodal content such as textual title, associated image, meta information (e.g., category, view count, and like count), and the popularity.

With the rapid advancement of multimedia technologies and the widespread use of social networks, people are increasingly inclined to share moments from their daily lives on social media platforms [27, 30, 37, 38]. However, the popularity of each post, which is measured by metrics such as view count and time since publication, can vary significantly due to factors like user preferences, follower count, and posting time [25, 26, 28, 34]. Accurately predicting a post's popularity is of great practical value. For example, businesses can identify content worth promoting through targeted advertising, while platforms can enhance user engagement by recommending popular posts and connections based on trending signals.

Despite its importance, predicting social media popularity is a challenging task. This is primarily due to the complex and heterogeneous nature of social media data, which includes both structured elements (e.g., user metadata) and unstructured content (e.g., images and texts) [16, 29, 33]. As shown in Figure 1, a typical post includes a blend of structured metadata (such as post category, total views, and follower count) and unstructured inputs like visual content and titles, making it difficult to handle all modalities effectively.

In recent years, an increasing number of studies have focused on this problem [2, 4, 5, 8, 11, 16, 21, 23, 31, 32]. For instance, HyFea [11] leverages CatBoost [19] to fuse multimodal features, while DFT-MOVLT [5] uses the ALBEF model [14] to combine metadata and text embeddings, applying contrastive learning [20] with image features to enhance representation learning. Although these methods have achieved promising results, they generally rely on unified models to process all input simultaneously. This design can underutilize important information, leading to suboptimal performance. For example, CatBoost-based models tend to prioritize structured statistical features (e.g., user ID, the length of post title) [8, 21], while in DFT-MOVLT, image features dominate the prediction.

To overcome these limitations, we introduce Fusion-Aware Multimodal Ensemble (FAME), a framework designed to better leverage the diverse information contained in social media posts. FAME integrates predictions from four complementary models. Drawing on the success of tree-based methods for both structured and some unstructured data [11, 21], we incorporate three strong tree-based learners—CatBoost [19], LightGBM [9], and AutoGluon [7]—to effectively handle both structured and unstructured data. In parallel, to better capture insights from unstructured modalities, we introduce a denoising autoencoder (DAE) [22]. The DAE is trained in two stages: first, it learns robust text and image representations by reconstructing noisy inputs; then, these representations are fused with original features to improve the prediction of post popularity. To further enhance representation learning, we incorporate multiple diverse textual and visual features into the FAME framework.

In summary, our contributions are:

- We propose the FAME framework for popularity prediction, which fuses the strengths of tree-based models for structured data and a DAE model for unstructured data.
- We introduce the DAE into social media popularity prediction to learn more robust unstructured representations.

2 Related Works

Predicting the popularity of social media posts involves estimating how widely a post will be viewed based on a variety of factors, including the user's profile (e.g., follower count), the content itself (text [15, 24, 35], images [29, 36], or videos [33, 34]), and additional metadata such as content category. Popularity is typically shaped by two main aspects: the attractiveness of the content, reflected by the number of views, and the recency of the post, as newer content tends to receive more attention [25, 28]. Accurate popularity prediction is essential for applications such as personalized content recommendation, friend suggestions, and targeted advertising. However, the task remains challenging due to the heterogeneous nature of the input data: structured metadata (e.g., follower count) is fundamentally different from unstructured content (e.g., text or images), making effective integration across modalities difficult.

Earlier research has addressed this challenge from a temporal perspective [26, 27, 30]. For example, the MT model [26] decomposes user-post-time interactions to capture temporal dynamics, while DTCN [27] incorporates attention-based temporal context learning across multiple time scales. MTD [30] further explores time-sensitive patterns through contextual matrix decomposition. These approaches emphasize temporal dependencies but often underutilize rich multimodal content. Other studies focus on user profiles and structured metadata, using tree-based models such as CatBoost [19], XGBoost [1], and LightGBM [9], which naturally highlight dominant structured features like follower count [11, 21]. More recent works attempt to align structured and unstructured modalities through unified embedding strategies. Methods like TTC-VLT [4], CL-WMTG [2], DFT-MOVLT [5], and DSPT [8] employ contrastive learning [14, 20] to enforce consistency between text, image, and metadata. While effective, these models often operate under a single framework, which may bias predictions toward specific modalities—e.g., profile metadata in tree-based models [8, 21] or visual features in contrastive learning approaches [5].

In contrast, the proposed FAME integrates outputs from multiple specialized models. By combining the strengths of structured-data learners and unstructured-data encoders, FAME offers a more balanced and effective solution for modeling the complex, multimodal nature of social media posts.

3 Methodology

As illustrated in Figure 2, the proposed FAME framework consists of three stages: Pre-Processing, Model Training, and Popularity Fusion. In the Pre-Processing stage, we extract and structure information from three sources associated with each social media post: (1) Meta information, such as the post's category, and the number of photos uploaded by the user; (2) Textual content, including post titles and post tags; and (3) Visual content, such as the uploaded image. These features are then passed to the Model Training stage, where four separate models, including CatBoost [19], LightGBM [9], AutoGluon [7], and DAE [22], independently predict the post's popularity. Finally, their outputs are aggregated in the Popularity Fusion stage using a weighted combination strategy to produce the final prediction. Each component is detailed below.

3.1 Pre-Processing Stage

In this section, we describe the feature extraction process for social media posts. Each post contains rich, multimodal content including text, images, and structured metadata (e.g., category, number of

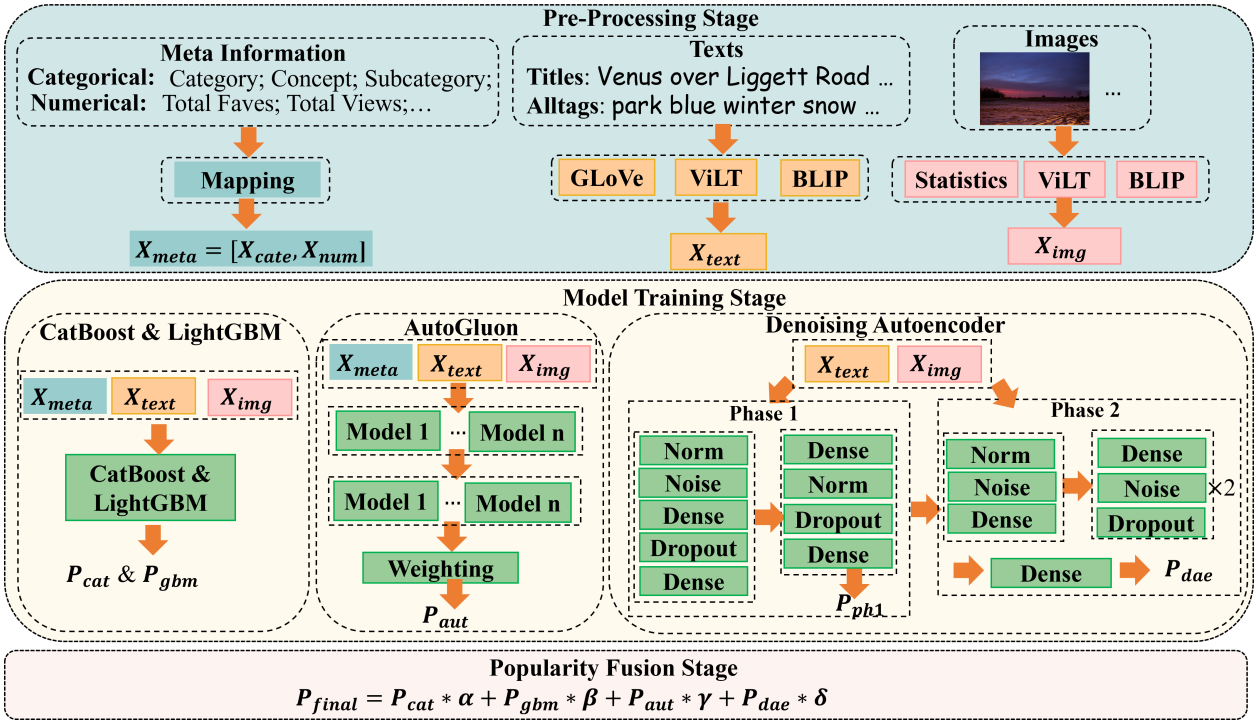

Figure 2: The overall framework of the proposed FAME model, consisting of three stages: a pre-processing stage, a model training stage, and a popularity fusion stage. In the first stage, multimodal features are extracted from the post's metadata, text, and image using tailored processing techniques. These features are then fed into four individual models—CatBoost, LightGBM, AutoGluon, and Denoising Autoencoder—to predict popularity scores. In the final stage, the predicted scores are fused through a weighted strategy to obtain the final popularity prediction.

views). As illustrated in Figure 1, we extract and process these inputs using three dedicated strategies.

3.1.1 Meta Information Processing. The meta information associated with a post contains both categorical and numerical attributes. For the categorical metadata, we extract three fields: category, concept, and subcategory. These are first mapped to numerical IDs, resulting in 11 distinct categories, 77 subcategories, and 668 concepts. We represent these encoded values as X_{cate}.

The remaining numerical metadata, including the number of views, likes, the length of the post title, is compiled into a vector X_{num}. We concatenate the categorical and numerical features to form the final meta representation: $X_{meta} = [X_{cate}, X_{num}]$.

3.1.2 Text and Image Processing. Inspired by recent work demonstrating that rich textual representations and alignment with visual features can significantly boost performance [8, 16, 21], FAME adopts efficient strategies for processing unstructured content. Unlike models such as DFT-MOVLT [5] that require computationally intensive contrastive learning to fine-tune pre-trained models[14], FAME directly leverages publicly available pre-trained models to extract feature representations. This makes our approach more practical and resource-friendly for real-world deployment.

To capture diverse multimodal features, three types of text and image representations are taken into consideration: (1) statistical features based on textual GloVe [17] embedding and image statistics, (2) embeddings from the ViLT model [10] pre-trained on Flickr30k [18], and (3) features from the BLIP model [12, 13].

GloVe Embeddings and Image Statistics. Following previous work [11], we use GloVe [17] embeddings to extract features from two textual components of each post: the title and all tags. These are individually encoded and reduced to 20 dimensions using SVD, then concatenated into a 40-dimensional textual feature, denoted as X_{text}. For image metadata, we extract basic statistics such as 'ImgLength', 'ImgWidth', 'Pixel count', and 'ImgModel' (which includes four types: P, L, RGB, and CMYK), resulting in a 4-dimensional image feature, denoted as X_{img}.

ViLT Embedding for texts and images. To better align textual and visual modalities, we use the ViLT model [10] pre-trained on Flickr30k [18] to extract unified embeddings for the post's title, all tags, and image. Each embedding is encoded into a 768-dimensional vector, facilitating multimodal fusion and comparison. Thus the dimensions for X_{text} and X_{img} are 1,536 and 768.

BLIP Embedding for texts and images. The BLIP model [12, 13], which is pre-trained on larger and more diverse datasets, enabling stronger semantic alignment between text and images, is used to obtain 768-dimensional embeddings for the title, all tags, and image, enhancing representational consistency across modalities. Thus the dimensions for X_{text} and X_{img} are 1,536 and 768.

3.2 Model Training Stage

In this stage, we introduce the models employed in our proposed FAME, including CatBoost, LightGBM, AutoGluon, and DAE.

3.2.1 CatBoost and LightGBM. Both CatBoost [19] and LightGBM [9] are gradient-boosted decision tree models. They automatically identify and utilize important features to capture complex, non-linear relationships. Here these two models take the concatenation of the post's metadata, textual, and visual features $X_{input} = [X_{meta}, X_{text}, X_{img}]$ as input and generate the final prediction P_{cat} and P_{gbm}.

3.2.2 AutoGluon. AutoGluon [7] framework handles structured and multimodal data efficiently. It automatically infers the data types of input features, such as categorical, integer, or float, and applies robust pre-processing techniques accordingly. As illustrated in Figure 2, the preprocessed inputs are fed into an ensemble of models, including CatBoost, LightGBM, and MultiModal Predictor. The MultiModal Predictor integrates textual, visual, and metadata inputs, leveraging the ELECTRA transformer [6] for effective multimodal fusion. The representations generated by these models are concatenated and passed through the ensemble again for a second-stage prediction. Finally, the outputs of each model are combined using a weighted average to produce the final popularity prediction P_{aut}.

3.2.3 Denoising Autoencoder. The Denoising Autoencoder is designed to enhance robustness by reconstructing input representations from their noisy embeddings [22]. As illustrated in Figure 2, the DAE operates in two phases. In the first phase, we inject Gaussian noise into the normalized input vector:

$$X_{noise} = \text{GaussianNoise}(\text{BatchNorm}(X_{input}); \sigma = 0.1). \quad (1)$$

Here $X_{input} = [X_{text}, X_{img}]$, which means meta information X_{meta} is not used in DAE. σ denotes the noise. This noisy input is then passed through a dense layer with ReLU activation:

$$encoded_{noise} = ReLU(W_1 X_{noise} + b_1). \quad (2)$$

Here, W_1 and b_1 are learnable parameters in the dense layer. Next, we apply dropout and reconstruct the input using another dense layer:

$$decoded = ReLU(W_2 \cdot \text{Dropout}(encoded_{noise}) + b_2). \quad (3)$$

Here, W_2 and b_2 are learnable parameters in the dense layer.To ensure that the learned representations capture popularity-related information, we use the reconstructed representation *decoded* to predict popularity via a 2-layer perceptron (MLP) with batch normalization and dropout:

$$P_{ph1} = W_4 \cdot \text{Dropout}(\text{BatchNorm}(ReLU(W_3 \cdot decoded + b_3))) + b_4. \quad (4)$$

Here W_3, W_4, b_1 and b_2 are learnable parameters in the MLP layer. The total training loss in this phase combines the reconstruction loss and the popularity prediction loss:

$$\mathcal{L}_{dae1} = \frac{1}{B} \sum_{i=1}^{B} (|P_i - P_{ph1,i}| + |decoded_i - X_{input,i}|). \quad (5)$$

Here B is the batch size, P_i is the ground-truth popularity of the i^{th} post. This design allows the DAE to generate more robust and informative representations for popularity prediction.

In the second training phase of the DAE, we reuse only the encoding layers from the first phase—specifically the layers responsible for generating the noisy representation $encoded_{noise}$ as described in Equations 1 and 2. These layers are kept frozen to preserve the robustness gained from the first phase. We then concatenate this robust representation with the original input features, forming a combined feature vector $X_{concat} = [encoded_{noise}, X_{input}]$.

This concatenated vector is passed through two sequential layers, each consisting of a dense layer followed by Gaussian noise injection and dropout:

$$X^1_{con} = \text{Dropout}(\text{GaussianNoise}(W_5 X_{concat} + b_5; \sigma = 0.1)), \quad (6)$$

and:

$$X^2_{con} = \text{Dropout}(\text{GaussianNoise}(W_6 X^1_{con} + b_6; \sigma = 0.1)). \quad (7)$$

Here W_5, W_6, b_5 and b_6 are learnable parameters, and σ is the noise. The resulting representation is used to predict the popularity score P_{dae} through a dense layer:

$$P_{dae} = W_7 X^2_{con} + b_7. \quad (8)$$

Here W_7 and b_7 are learnable parameters. The second phase is trained using the MAE between the predicted and ground truth popularity values:

$$\mathcal{L}_{dae2} = \frac{1}{B} \sum_{i=1}^{B} |P_i - P_{dae,i}|. \quad (9)$$

3.3 Popularity Fusion Stage

In this stage, we combine the predictions from the four models introduced in the Model Training Stage to generate a final, more comprehensive popularity score. This is achieved through a weighted ensemble of the individual model outputs:

$$P_{final} = P_{cat} \times \alpha + P_{gbm} \times \beta + P_{aut} \times \gamma + P_{dae} \times \delta. \quad (10)$$

Here, α, β, γ and δ are hyper-parameters that control the contribution of each model's prediction, subject to the constraint:

$$\alpha + \beta + \gamma + \delta = 1.0. \quad (11)$$

This fusion strategy enables the model to leverage complementary strengths of different models and produce more robust and accurate predictions.

4 Experiments

4.1 Dataset and Evaluation Metrics

4.1.1 Dataset. We evaluate our proposed FAME framework on the Social Media Prediction Dataset (SMPD) [25, 28, 29], which contains posts collected from the Flickr platform over a 480-day period, from November 2015 to March 2016. The dataset is split chronologically into a training set and a testing set. The training set includes 305,613 posts from over 38k users, with an average popularity score of 6.41 and a standard deviation of 2.47. The test set includes 180,581 posts from more than 31k users, with an average popularity score of 5.12 and a standard deviation of 2.41. Each post contains multiple modalities: textual data (e.g., title and tags), visual

data (image), and metadata related to the post and its author (e.g., post category, total number of photos uploaded by the user, total views across all of the user's posts, etc.).

4.1.2 Evaluation Metrics. Following prior work [3, 5, 11, 21], we use Spearman's Rank Correlation Coefficient (SRC) and Mean Absolute Error (MAE) as our evaluation metrics. SRC measures the monotonic relationship between predicted and ground-truth popularity scores by assessing the rank correlation between them using the following equation:

$$SRC = \frac{1}{N-1}\sum_{i=1}^{N}(\frac{P_i-\bar{P}}{\sigma_P})(\frac{\hat{P}_i-\bar{\hat{P}}}{\sigma_{\hat{P}}}). \tag{12}$$

where P_i and $\hat{P}_i$ are the ground-truth and predicted popularity scores. $\bar{P}$, σ_P and $\bar{\hat{P}}$, $\sigma_{\hat{P}}$ are their respective means and standard deviations. N is the number of the samples.

MAE quantifies the average absolute difference between predicted and ground-truth popularity:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|\hat{P}_i - P_i|. \tag{13}$$

4.2 Implementation Details

All experiments are conducted on two NVIDIA RTX 3090 GPUs with CUDA version 12.2. For DAE model, we use a batch size of 256 and train for 100 epochs in the first phase, with early stopping set to 10 epochs and a learning rate of 0.001. In the second phase, we perform five-fold cross-validation with a batch size of 2048, training for up to 200 epochs, using early stopping after 8 epochs, and maintaining a learning rate of 0.001. For CatBoost and LightGBM models, we use a learning rate of 0.03, a tree depth of 8, and run up to 200,000 iterations. For AutoGluon, we use a batch size of 128, a learning rate of 0.0001, and train for 10 epochs. And the hyper-parameters of the popularity fusion stage are: $\alpha = 0.4$, $\beta = 0.35$,$\gamma = 0.15$, $\delta = 0.1$.

4.3 Comparison to State-of-the-Art Methods

4.3.1 Baselines. We compare FAME with seven state-of-the-art social media popularity prediction approaches: CL-WMTG [2], HyFea [11], TTC-VLT [4], DEVL [31], HVLA [21], TTME [3], and DFT-MOVLT [5].

CL-WMTG[2] incorporates object-level relational modeling within images using graph learning and fuses visual-textual information via curriculum learning.

HyFea[11] leverages CatBoost to combine text, image, and metadata features in a unified prediction framework.

TTC-VLT[4] proposes a title-tag contrastive learning strategy to distinguish visual associations across different textual modalities.

DEVL[31] extracts fine-grained visual and textual semantics, using multi-perspective image parsing and hierarchical language embeddings to assess post attractiveness.

HVLA[21] aligns visual and language features in a shared semantic space by using BLIP-2 to generate image descriptions and extract embeddings from multimodal inputs.

TTME[3] introduces AI-generated content and a mixture of modality experts in the pre-training phase to better handle noisy data and tri-modal learning, further employing knowledge distillation for robust adaptation.

DFT-MOVLT[5] enhances vision-language pre-training by introducing compound text (numerical + textual data) and applying multi-objective learning, followed by dual fine-tuning strategies for generalized prediction.

Table 1: Experimental results of FAME and baselines. Best performance is bold.

Models	SRC (↑)	MAE (↓)
CL-WMTG [2]	0.6545	1.3931
HyFea [11]	0.702	1.383
TTC-VLT [4]	0.7118	1.2691
DEVL [31]	0.7149	1.2371
HVLA [21]	0.7341	1.1970
TTME [3]	0.7434	1.1745
DFT-MOVLT [5]	0.753	1.170
FAME (Ours)	**0.7543**	**1.1607**

Table 2: Ablation studies of each model in FAME. Best performance is bold.

Models	SRC (↑)	MAE (↓)
FAME(full model)	**0.7543**	**1.1607**
w/o CatBoost	0.7360	1.2091
w/o LightGBM	0.7461	1.1845
w/o AutoGluon	0.7475	1.1809
w/o DAE	0.7511	1.1711

4.3.2 Performance Comparison. As shown in Table 1, FAME consistently outperforms all baselines on the SMPD dataset in terms of both SRC and MAE, achieving new state-of-the-art results. These improvements highlight FAME's stronger ability to accurately capture both the absolute and relative popularity of social media posts. Notably, models that include visual-textual alignment techniques, such as TTC-VLT, HVLA, and DFT-MOVLT, also show competitive performance, but still fall short of FAME. One possible reason is that these methods focus mainly on aligning image and text modalities, while neglecting discrete metadata such as post category or author information. These metadata elements often differ significantly in distribution and representation from visual-textual features, and naïvely combining them may introduce noise.

In contrast, FAME employs tree-based models for both structured metadata and unstructured text and images, and adapts a DAE for unstructured text and image inputs. The final prediction is obtained by fusing the outputs from each model, effectively leverage complementary strengths of different models and produce more robust and accurate predictions.

4.4 Ablation Studies

To assess the individual contributions of each model within the FAME framework, we conduct a series of ablation experiments, as reported in Table 2. Removing the DAE leads to the smallest performance decline, with the SRC decreasing by 0.0032 and the MAE increasing by 0.0104. This suggests that while the DAE enhances

Table 3: Experimental results with various settings. Best performance is bold.

meta	text			image			DAE		LightGBM		CatBoost	
	GLoVe	ViLT	BLIP	Statistics	ViLT	BLIP	SRC (↑)	MAE (↓)	SRC (↑)	MAE (↓)	SRC (↑)	MAE (↓)
	✓			✓			0.6488	1.6727	-	-	-	-
		✓			✓		0.6267	1.4273	-	-	-	-
			✓			✓	0.6735	1.3723	-	-	-	-
✓	✓			✓			0.6107	1.4734	0.7283	1.2479	0.7311	1.2307
✓		✓			✓		0.5957	1.5138	0.7040	1.2895	0.7184	1.2615
✓			✓			✓	0.6267	1.4273	0.7313	1.2241	0.7322	1.2215
✓	✓		✓	✓		✓	0.6349	1.4131	0.7337	1.2317	**0.7343**	**1.2152**

the robustness of unstructured feature representations, its contribution is more supplementary. In contrast, excluding CatBoost results in the most substantial performance drop: SRC decreases by 0.0183 and MAE increases by 0.0484, which highlights its critical role in modeling structured metadata. The removal of LightGBM and AutoGluon causes moderate degradation in performance, falling between the above two cases. These results are consistent with previous studies [21], which emphasize CatBoost's strong ability to handle mixed-type data, especially in datasets like SMPD that contain both discrete and continuous attributes.

4.5 Quantitive Analysis

We further investigate how different combinations of data modalities and feature representations affect each individual model's performance. Table 3 presents detailed results for DAE, CatBoost, and LightGBM under various input settings. For DAE, we find that using BLIP-derived image and text embeddings yields the best performance. Interestingly, when using GloVe for text and simple image statistics, performance surpasses more complex ViLT-based features, possibly due to better generalizability.

However, adding metadata to the DAE (i.e., assigning $X_{input} = [X_{meta}, X_{text}, X_{img}]$ consistently degrades performance. This may stem from the fact that metadata contains mixed types—e.g., categories, and large numeric values like "957, 971 total views" or "149 followers" in Figure 1—which can introduce significant variance. These inputs, when treated as continuous features and fed into the same encoder as unstructured data, can amplify small noise and disrupt learning.

CatBoost and LightGBM show similar trends. BLIP-based features outperform ViLT-derived ones, and combinations using GloVe text features and basic image statistics often yield the best results. These results suggest that careful feature engineering and modality-specific processing are essential for robust popularity prediction.

4.6 Prediction Analysis

To better understand the behavior of the four models used in FAME, we visualize the predicted popularity distributions of CatBoost, LightGBM, AutoGluon, and DAE on the test set of the SMPD dataset, which is shown in Figure 3. We observe that the tree-based models produce similar prediction trends: the popularity density first rises and then falls, peaking within the range of 3.55 to 4.78. In contrast, DAE, which relies solely on textual and visual features, shows a similar rising-falling pattern, but with a peak shifted to a higher

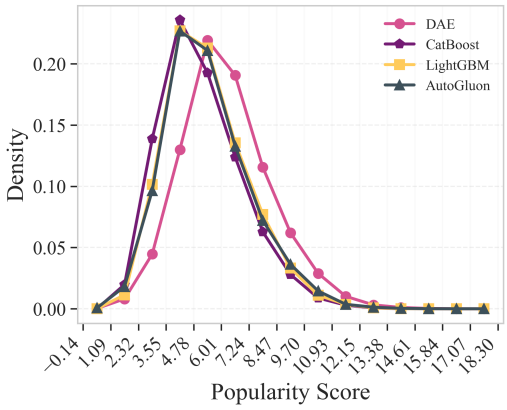

Figure 3: Predicted popularity density distributions on the SMPD test set for the four components of FAME.

range of 4.78 to 6.01. Additionally, the DAE model tends to predict higher popularity values overall compared to the tree-based models. This discrepancy may arise because the tree-based models can better leverage structured metadata—such as the user's overall view count across all posts—which allows for more grounded predictions. These results highlight the distinct predictive tendencies and biases of each model. By fusing the outputs of these diverse models, FAME captures a broader range of patterns and ultimately achieves more accurate and robust predictions.

5 Conclusion

This paper introduces FAME, a Fusion-Aware Multimodal Ensemble framework for predicting social media popularity from heterogeneous inputs, including text, images, and structured metadata. To better capture unstructured content, FAME integrates a DAE trained on textual and visual features, mitigating the dominance of structured data. Predictions from three tree-based models and the DAE are fused using a weighted strategy, leading to more accurate and balanced results. Experiments on the SMPD dataset demonstrate that FAME achieves state-of-the-art performance, ranking first in the Image Track of the SMP Challenge 2025. While FAME delivers strong results, its multi-model nature increases computational cost. Future work will explore more efficient fusion strategies to maintain accuracy with lower computation overhead.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No.U24A20250 and No.62176048), the Natural Science Foundation of Sichuan Province (Grant No.2024NSFTD0042, No.2024YFG0006, No.2024NSFSC0506, and No.2025ZNSFSC1487), and the Fundamental Research Funds for the Central Universities (No.ZYGX2024J022 and No.ZYGX2024Z005).

References

[1] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 785–794.

[2] Weilong Chen, Feng Hong, Chenghao Huang, Shaoliang Zhang, Rui Wang, Ruobing Xie, Feng Xia, Leyu Lin, Yanru Zhang, and Yan Wang. 2020. Curriculum learning for wide multimedia-based transformer with graph target detection. In *Proceedings of the 28th ACM International Conference on Multimedia*. 4575–4579.

[3] Weilong Chen, Wenhao Hu, Xiaolu Chen, Weimin Yuan, Yan Wang, Yanru Zhang, and Zhu Han. 2024. Tri-Modal Transformers with Mixture-of-Modality-Experts for Social Media Prediction. *IEEE Transactions on Circuits and Systems for Video Technology* (2024).

[4] Weilong Chen, Chenghao Huang, Weimin Yuan, Xiaolu Chen, Wenhao Hu, Xinran Zhang, and Yanru Zhang. 2022. And-tag contrastive vision-and-language transformer for social media popularity prediction. In *Proceedings of the 30th ACM International Conference on Multimedia*. 7008–7012.

[5] Xiaolu Chen, Weilong Chen, Chenghao Huang, Zhongjian Zhang, Lixin Duan, and Yanru Zhang. 2023. Double-Fine-Tuning Multi-Objective Vision-and-Language Transformer for Social Media Popularity Prediction. In *Proceedings of the 31st ACM International Conference on Multimedia*. 9462–9466.

[6] Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *International Conference on Learning Representations*.

[7] Nick Erickson, Jonas Mueller, Alexander Shirkov, Hang Zhang, Pedro Larroy, Mu Li, and Alexander Smola. 2020. Autogluon-tabular: Robust and accurate automl for structured data. *arXiv preprint arXiv:2003.06505* (2020).

[8] Wenhao Hu, Weilong Chen, Weimin Yuan, Yan Wang, Shimin Cai, and Yanru Zhang. 2024. Dual-Stream Pre-Training Transformer to Enhance Multimodal Learning for Social Media Prediction. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 11450–11456.

[9] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems* 30 (2017).

[10] Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*. PMLR, 5583–5594.

[11] Xin Lai, Yihong Zhang, and Wei Zhang. 2020. HyFea: Winning solution to social media popularity prediction for multimedia grand challenge 2020. In *Proceedings of the 28th ACM International Conference on Multimedia*. 4565–4569.

[12] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*. PMLR, 19730–19742.

[13] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*. PMLR, 12888–12900.

[14] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems* 34 (2021), 9694–9705.

[15] Wei Li, Jiawen Deng, Jiali You, Yuanyuan He, Yan Zhuang, and Fuji Ren. 2025. ETS-MM: A Multi-Modal Social Bot Detection Model Based on Enhanced Textual Semantic Representation. In *Proceedings of the ACM on Web Conference 2025*. 4160–4170.

[16] Yu-Shi Lin and Anthony JT Lee. 2024. MMF: Winning Solution to Social Media Popularity Prediction Challenge 2024. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 11445–11449.

[17] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[18] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*. 2641–2649.

[19] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. 2018. CatBoost: unbiased boosting with categorical features. *Advances in neural information processing systems* 31 (2018).

[20] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PmLR, 8748–8763.

[21] Mingsheng Tu, Tianjiao Wan*, Qisheng Xu, Xinhao Jiang, Kele Xu, and Cheng Yang. 2024. Higher-Order Vision-Language Alignment for Social Media Prediction. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 11457–11463.

[22] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. 2008. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*. 1096–1103.

[23] Kai Wang, Penghui Wang, Xin Chen, Qiushi Huang, Zhendong Mao, and Yongdong Zhang. 2020. A feature generalization framework for social media popularity prediction. In *proceedings of the 28th ACM International Conference on Multimedia*. 4570–4574.

[24] Bai Wei and Zhuang Yan. 2021. Claim Stance Classification Optimized by Data Augment. In *2021 18th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*. IEEE, 171–174.

[25] Bo Wu, Wen-Huang Cheng, Peiye Liu, Bei Liu, Zhaoyang Zeng, and Jiebo Luo. 2019. Smp challenge: An overview of social media prediction challenge 2019. In *Proceedings of the 27th ACM International Conference on Multimedia*. 2667–2671.

[26] Bo Wu, Wen-Huang Cheng, Yongdong Zhang, and Tao Mei. 2016. Time matters: Multi-scale temporalization of social media popularity. In *Proceedings of the 24th ACM international conference on Multimedia*.

[27] Bo Wu, Wen-Huang Cheng, Yongdong Zhang, Huang Qiushi, Li Jintao, and Tao Mei. 2017. Sequential Prediction of Social Media Popularity with Deep Temporal Context Networks. In *International Joint Conference on Artificial Intelligence (IJCAI)*.

[28] Bo Wu, Peiye Liu, Wen-Huang Cheng, Bei Liu, Zhaoyang Zeng, Jia Wang, Qiushi Huang, and Jiebo Luo. 2023. SMP Challenge: An overview and analysis of social media prediction challenge. In *Proceedings of the 31st ACM International Conference on Multimedia*. 9651–9655.

[29] Bo Wu, Peiye Liu, Qiushi Huang, Zhaoyang Zeng, Jia Wang, Bei Liu, Jiebo Luo, and Wen-Huang Cheng. 2024. SMP Challenge Summary: Social Media Prediction Challenge. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 11442–11444.

[30] Bo Wu, Tao Mei, Wen-Huang Cheng, and Yongdong Zhang. 2016. Unfolding Temporal Dynamics: Predicting Social Media Popularity Using Multi-scale Temporal Decomposition. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI)*.

[31] Jianmin Wu, Liming Zhao, Dangwei Li, Chen-Wei Xie, Siyang Sun, and Yun Zheng. 2022. Deeply exploit visual and language information for social media popularity prediction. In *Proceedings of the 30th ACM International Conference on Multimedia*. 7045–7049.

[32] Kele Xu, Zhimin Lin, Jianqiao Zhao, Peicang Shi, Wei Deng, and Huaimin Wang. 2020. Multimodal deep learning for social media popularity prediction with attention mechanism. In *Proceedings of the 28th ACM International Conference on Multimedia*. 4580–4584.

[33] Xovee Xu, Yifan Zhang, Fan Zhou, and Jingkuan Song. 2025. Improving Multimodal Social Media Popularity Prediction via Selective Retrieval Knowledge Augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 932–940.

[34] Yijie Xu, Bolun Zheng, Wei Zhu, Hangjia Pan, Yuchen Yao, Ning Xu, Anan Liu, Quan Zhang, and Chenggang Yan. 2025. SMTPD: A New Benchmark for Temporal Prediction of Social Media Popularity. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 18847–18857.

[35] Yan Zhuang and Yanru Zhang. 2022. Yet at Factify 2022: Unimodal and bimodal roberta-based models for fact checking. In *Proceedings of De-Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection, CEUR*.

[36] Yan Zhuang and Yanru Zhang. 2022. Yet at Memotion 2.0 2022: Hate speech detection combining bilstm and fully connected layers. In *Proceedings of De-Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection, CEUR*.

[37] Yan Zhuang, Yanru Zhang, Jiawen Deng, and Fuji Ren. 2025. R3DG: Retrieve, Rank and Reconstruction with Different Granularities for Multimodal Sentiment Analysis. *Research* (2025).

[38] Yan Zhuang, Yanru Zhang, Zheng Hu, Xiaoyue Zhang, Jiawen Deng, and Fuji Ren. 2024. GLoMo: Global-Local Modal Fusion for Multimodal Sentiment Analysis. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 1800–1809.